Research Report
ETS RR–15-03

Evaluation of *e-rater*® for the *Praxis I*® Writing Test

Chaitanya Ramineni

Catherine S. Trapani

David M. Williamson

June 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Evaluation of *e-rater*® for the *Praxis I*® Writing Test

Chaitanya Ramineni, Catherine S. Trapani, & David M. Williamson

Educational Testing Service, NJ

Automated scoring models were trained and evaluated for the essay task in the *Praxis I*® writing test. Prompt-specific and generic *e-rater*® scoring models were built, and evaluation statistics, such as quadratic weighted kappa, Pearson correlation, and standardized differences in mean scores, were examined to evaluate the e-rater model performance against human scores. Performance of the scoring model was also evaluated across different demographic subgroups using the same statistics. Additionally, correlations for automated scores with external measures were observed for validity evidence. Analyses were performed to establish appropriate agreement thresholds between human and e-rater scores for unusual essays and to examine the impact of using e-rater on operational scores and classification rates. The generic e-rater scoring model was recommended for operational use to produce contributory scores within a discrepancy threshold of 1.5 with a human score.

Keywords Automated scoring; e-rater; PRAXIS writing test; generic scoring model; contributory scoring

doi:10.1002/ets2.12047

The *Praxis I*® *PPST*® (Educational Testing Service [ETS], 2014) assessment is a teacher licensure test that measures basic reading, writing, and mathematics skills of candidates preparing for a career in teaching. Multiple states across the United States require Praxis I test scores as part of their teacher-licensing process; some colleges and universities use the Praxis I test scores for admissions decisions for their teacher education programs. The Praxis I tests are delivered mostly online year-round by appointment, while paper-delivered tests are also available on prescheduled dates. States and agencies that use the *Praxis* test scores for licensing decisions set their own requirements and passing scores.

The Praxis I PPST consists of three separate tests:

- a reading test to measure the ability to understand, analyze, and evaluate written messages;
- a mathematics test to measure mathematical skills and concepts that an educated adult might need; and
- a writing test to assess the ability to use grammar and language appropriately and to communicate effectively in writing.

Praxis I tests are generally multiple-choice (MC) tests, but the writing test also includes an essay section. The MC section of the writing test has multiple items assessing grammatical relationships, structural relationships, and word choice and mechanics, and the essay section has one essay task evaluating general writing ability through organization, development, language control, and effective sentence construction. The essay task is timed for 30 minutes and scored using a 6-point holistic rubric. The prompts for the essay task are created by test developers at ETS using rigorous standardized procedures (see Baldwin, Fowles, & Livingston, 2005), which include both content review and fairness review of prompts by ETS experts, as well as content advisory groups.

Background

With the rise in use of constructed response (CR) items within the last decade, many other high-stakes assessments, such as the *GRE*®, the *SAT*®, and the Graduate Management Admission Test (GMAT), have included CR items in speaking and/or writing sections. Some of these tests include more than one CR item in the same test; in fact, the new Praxis Core Academic Skills for Educators (Core) tests will contain two CR tasks as part of the writing test. CR items are believed to measure aspects of a construct that are not adequately addressed through MC items. However, compared to their MC counterparts, such items take longer to administer with smaller contributions to reliability per unit time, and they delay

Corresponding author: C. Ramineni, E-mail: cramineni@ets.org

score reporting due to the additional effort and expense typically required to recruit, train, and monitor human raters. With the challenges for efficient scoring and continuously rising interest in use of CR items, there is potential value for using automated means to score CR tasks to either augment or replace human scorers.

Automated scoring systems, in particular systems designed to score a particular type of response that is in relatively widespread use across various assessments, purposes, and populations, can provide a greater degree of scoring efficiency and complementary construct coverage alongside human raters. Examples of automated scoring systems include essay scoring systems (Shermis & Burstein, 2003), automated scoring of mathematical equations (Risse, 2007; Singley & Bennett, 1998), scoring short written responses for correct answers to prompts (Callear, Jerrams-Smith, & Soh, 2001; Leacock & Chodorow, 2003; Mitchell, Russell, Broomhead, & Aldridge, 2002; Sargeant, Wood, & Anderson, 2004; Sukkarieh & Pulman, 2005), and the automated scoring of spoken responses (Bernstein, De Jong, Pisoni, & Townshend, 2000; Chevalier, 2007; Franco et al., 2000; Xi, Higgins, Zechner, & Williamson, 2008; Zechner & Bejar, 2006). Of these, automated scoring technologies for the traditional essay responses are most popular, with more than 12 different automated essay evaluation systems available for scoring and/or for performance feedback and improvement of writing quality. The most widely known of these systems include the Knowledge Analysis Technologies (KAT) engine 5 (Landauer, Laham, & Foltz, 2003), *e-rater*® (Attali & Burstein, 2006; Burstein, 2003), Project Essay Grade (Page, 1966, 1968, 2003), and IntelliMetric (Rudner, Garcia, & Welch, 2006). Each of these engines targets a generalizable approach of modeling or predicting human scores, yet each takes a somewhat different approach to achieving the desired scoring, both through different statistical methods and through different formulations of what features of writing are measured and used in determining the score. An explanation of how these systems work is beyond the scope of this paper, except for the e-rater system used for the analyses in this report, for which such an explanation is provided below.

Automated scoring, in general, can provide some advantages similar to that of MC scoring, including fast scoring, constant availability of scoring, lower per-unit costs, reduced coordination efforts for human raters, greater score consistency, and a higher degree of tractability of score logic for a given response. Additionally, automated scoring also offers the potential for providing a degree of performance-specific feedback that is not feasible under operational human scoring for CR tasks. These advantages facilitate increased use of CR items for large-scale assessments. However, accompanying such potential advantages is a need to evaluate the cost and effort of developing such systems and the potential for vulnerability in scoring unusual or bad-faith responses inappropriately, to validate the use of such systems, and to critically review the construct that is represented in resultant scores.

The automated scoring models based on e-rater have been successfully evaluated in recent years for the writing prompts included in the old GRE General Test (Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012a), the *TOEFL iBT*® test (Attali, Bridgeman, & Trapani, 2010; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012b), and the GRE revised General Test (Breyer et al., 2014). The current TOEFL iBT test uses e-rater for operational scoring of the essay tasks, and the GRE uses e-rater as a quality control on the reported human scores, thus allowing the programs to report scores efficiently and to use their human rater pool more effectively.

Motivation

The *TOEFL*® independent writing task and GRE issue writing task require the examinees to support an opinion similar to the Praxis I essay topics that require examinees to draw from personal experience, observation, or reading to support a position using reasons and examples without specialized knowledge of a topic. Also, similar to the scoring rubrics for the TOEFL and the GRE tasks, the responses to the Praxis essay topics are scored holistically on aspects related to clarity, organization, and development, and use of language conventions.

The successful evaluation and deployment of automated scoring models for scoring the TOEFL and the GRE prompts, along with the projected cost and time benefits for operational use, supported evaluation of e-rater for scoring the Praxis I essay task. Hence, the purpose of this study was to develop and evaluate e-rater automated scoring models for the CR items in the Praxis I writing test. In particular, in this study, we investigated if an e-rater score could successfully replace the score of one of the two human raters in operational scoring of CR items in the Praxis I writing test, thereby effectively reducing the program costs and ensuring fast and consistent score turnaround for the large number of test takers and prospective teaching licensure applicants who take the test year-round at several computer-based test centers. The responses from the paper-based tests are not readily conducive to automated scoring model building and evaluation work and require

additional transcription effort. Computer-based tests that produce electronic essay responses as part of the test are readily usable for automated scoring work and were, therefore, used as the source of data for this study.

Scoring Rules for the Praxis I Writing Test

The Praxis I writing test contains 38 operational MC items in three categories assessing grammatical relationships, structural relationships, and word choice and mechanics. The possible raw score range for the MC items is 0 – 38. The assessment also includes one essay item scored by two raters on a scale of 1 – 6 evaluating general writing ability. The ratings are added on the individual item, without any weighting, to yield a possible raw score range of 2 – 12 for the CR item. If the two human ratings differ by more than 1 point, the essay response is sent for adjudication. The adjudicator's rating is combined with the initial human rating that is closest to it to determine the final combined score. If both the initial ratings are equidistant from the adjudicators' score, then the final combined CR score is given as 2 times the adjudicator score. The MC categories have a weight of 1, whereas the final combined score for the CR category is weighted by 3.1667 to increase its maximum contribution to the total score to 38 points, which is 50% of the total writing test score. The raw scores for MC and weighted CR (converted to an integer) are added to produce the final raw score, which is then converted to a scaled score of 150 – 190 and reported in 1-point increments on the scale.

Automated Scoring With e-rater

The e-rater is a computer program that scores essays primarily on the basis of features that are related to writing quality. The initial version of e-rater (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998) used more than 60 microfeatures to assess quality of writing in written assessments. In e-rater v2 (Attali & Burstein, 2006), the features were combined into a smaller set of macrofeatures related to general dimensions of writing quality. This set of features is constantly refined and enhanced in newer versions of e-rater, with e-rater v13.1 in operation at the time of writing this report. The e-rater program essentially uses natural language processing technology to evaluate a number of characteristics of the essay, including grammar, usage, mechanics, and development. These characteristics of essay quality are used to derive a prediction of the human score for the same response by using a statistical modeling procedure.

Features

The version 13.1 of e-rater (v13.1) uses 12 features, with 10 representing aspects of writing quality and two representing content. Most of these primary scoring features are formed by aggregating several microfeatures. For example, the mechanics feature is formed by aggregating microfeatures such as spelling errors, hyphen error, and compound word. An illustration of the construct decomposition of e-rater resulting from this structure is provided in Figure A1, where the features encapsulated in bold are entered as the predictor variables in the scoring model, and the other features are an incomplete illustrative listing of subfeatures measuring aspects of writing quality. The scoring features of e-rater are mapped to the six-trait model (Culham, 2003) commonly used to evaluate writing by teachers as described by Quinlan, Higgins, and Wolff (2009). More information on the microfeatures is available in Ramineni et al. (2012a).

Grammar, usage, mechanics, and style together identify more than 30 error types, including errors in subject-verb agreement, homophone errors, misspelling, and overuse of vocabulary. These are summarized for each feature as proportions of error rates relative to the essay length. Organization and development features are based on automatically identifying sentences in an essay as they correspond to essay-discourse categories: introductory material (background), thesis, main ideas, supporting ideas, and conclusion. For the organization feature, e-rater identifies the number of elements present for each category of discourse in an essay. For the development feature, e-rater computes the average length for all the discourse elements (in words) in an essay. Lexical complexity of the essay is represented by two features. The first is computed through a word frequency index used to obtain a measure of vocabulary level. The second feature computes average word length across all words in the essay and uses this as an index of sophistication of vocabulary usage. A combined feature indicative of correct use of collocations and prepositions in the essay is included as a measure of positive attributes of writing. An additional stylistic feature positively correlated with the human score, and evaluating the use of sentence variety by the writer, was also added to e-rater more recently.

Two prompt-specific vocabulary usage features relate to content of vocabulary used in the essay. Both features are based on the tendency to use words typical of those used in prior essays. The first feature indicates the score-point level to which the essay text is most similar with regard to vocabulary usage. The second analyzes the similarity of essay vocabulary to prior essays with the highest score point on the scale. These features are only used when building scoring models customized to individual prompts.

The e-rater Scoring Models

Developing e-rater scoring models is typically a fully automated and standardized two-stage process consisting of (a) model training/building and (b) model evaluation for which available data are randomly split into two subsets, referred to as model build (training) and cross-validation (evaluation) sets. Because the feature weights are estimated empirically so as to maximize agreement with human scores, any evaluation based on the training sample will tend to overstate a scoring model's performance. Thus, a more appropriate measure of performance can be obtained by applying the model to the independent evaluation sample. Subsequently, the feature scores and weights are applied to samples of essays in the evaluation set to produce an overall e-rater score and validate the model performance. In general, model performance will appear slightly degraded in this sample in comparison to the training sample. Models are evaluated and recommended for operational use if the results of automated scoring are comparable with agreement between two human raters on the evaluation sample. The quality of the estimated e-rater scoring models, and the effective functioning of the models in operational settings, depends critically on the nature and quality of the training and evaluation data. Therefore, certain standards that are used to guide the collection and analyses of the data for building and evaluation of automated scoring models have been developed by the automated scoring group at ETS (Ramineni & Williamson, 2013; Williamson, Xi, & Breyer, 2012). These include choosing a representative sample of double-scored essays in electronic format and a sufficient number of prompts and minimum test-taker sample sizes for model building. For the standards that are only partially or not met, there are subsequent implications when interpreting the results.

Prior to model build, the selected essay set is subjected to advisory flag analyses. Advisory flags act as filters and mark unusual characteristics because of which an essay would be identified as inappropriate for automated scoring. Each advisory flag marks a different problem. The use of these flags for an assessment is evaluated by comparing when e-rater considers an essay inappropriate versus when a human rater considers an essay inappropriate or off-topic. All advisories are evaluated individually, as well as combined. That is, individual advisories for which e-rater is found to effectively (on par with humans) identify essays that are inappropriate for automated scoring are combined sequentially and subjected to a similar evaluation against human markings. This process of advisory flag analyses helps determine which group of advisories aid e-rater in effectively screening for inappropriate essays and should be included as part of the operational e-rater framework for an assessment. Subjecting the sample of essays to advisory flagging prior to model build improves quality of model build by filtering the inappropriate essays from going into the model build phase for e-rater.

Advisory flags for e-rater are coded depending on the type(s) of issue(s) identified. Table 1 lists the names, a brief description, and numerical codes for all the advisory flags. An essay can be flagged for single or multiple issues. For instance, if an essay contains repetition of words, the flag will be set to 2 (reuse of language). However, if an essay contains repetition of words and is not relevant to the assigned topic, the flag will be set to 10, that is, 2 (reuse of language) +8 (not relevant). Flags 64 (too brief) and higher are referred to as fatal flags, where e-rater does not produce a score for the essay response, while the other flags trigger warnings along with the e-rater score.

Upon excluding the responses with fatal advisory flags, the e-rater engine evaluates the writing characteristics of the essays in the model build set to produce feature scores. Subsequently, weights for features are derived by entering them as predictor variables in a multiple linear regression modeling procedure (e.g., Cohen, Cohen, West, & Aiken, 2003) where holistic human ratings are the predicted criterion variable. These feature weights can then be applied to additional essays in the cross-validation set to produce a predicted score.

Two different types of e-rater scoring models are common and were built for this study: *prompt-specific models* and *generic models*. Following is a brief description of each type of model:

- Prompt-specific models. These models are custom-built for each prompt in the item pool. They are designed to provide the best-fit models for the particular prompt in question, with regression parameters scaled to match the

Table 1 Advisory Flag Code, Name, and Description

Advisory flag code	Flag name	Flag description
2	Reuse of language	Compared to other essays written on this topic, the essay contains more reuse of language, a possible indication that it contains sentences or paragraphs that are repeated.
4	Key concepts	Compared to other essays written on this topic, the essay shows less development of the key concepts on this topic.
8	Not relevant	The essay might not be relevant to the assigned topic.
16	Restatement	The essay appears to be a restatement of the topic with few additional concepts.
32	No resemblance	The essay does not resemble others that have been written on this topic, a possible indication that it is about something else or is not relevant to the issues the topic raises.
64	Too brief	The essay is too brief to evaluate.
128	Excessive length	The essay is longer than essays that can be accurately scored and must be within the word limit to receive a score.
256	Unidentifiable organizational elements	The essay could not be scored because some of its organizational elements could not be identified.
512	Excessive number of problems	The essay could not be scored because too many problems in grammar, usage, mechanics, and style were identified.
1024[a]	Unexpected topic	The essay appears to be on a subject that is different from the assigned topic.
2048[a]	Nonessay	The text submitted does not appear to be an essay.

[a]Not applicable for the Praxis I program.

characteristics of the human score distribution for the prompt. Prompt-specific models also include the two features for measuring use of prompt-specific vocabulary.

- Generic models. The smaller set of features derived in e-rater v2 enabled use of a single scoring model, referred to as a generic model, and the same feature weights across all prompts of an assessment. Generic models are based on taking a group of related prompts, typically 10 or more, and calibrating a regression model across all prompts so that the resultant model is the best fit for predicting human scores across all the prompts. As such, a common set of feature weights and a single intercept are used for all prompts, regardless of the particular prompt in the set. Generic models do not include features related to prompt-specific vocabulary and, thus, address only general writing quality independent of the topic. The generic modeling approach has the advantage of requiring smaller test-taker sample sizes per prompt (with enough prompts) and providing a uniform set of scoring criteria, regardless of the prompt delivered operationally. While prompt-specific models are the best-fit models for a given prompt, a generic model is preferred for the ease of implementation and maintenance.

The generic with prompt-specific intercept is a variant of the generic model that offers a common set of feature weights for all features with a customized regression intercept for each prompt but is less preferred due to substantial computational load and potentially marginal gains.

Evaluation Criteria

After the automated (e-rater) scores for all essays have been calculated, ETS uses professional evaluation standards and quality-control criteria to assess the performance of the models (see, e.g., Ramineni & Williamson, 2013; Williamson et al., 2012). Flagging conditions or thresholds are attached to the evaluation statistics to serve as warnings of potential performance problems. However, the flags are used as guidelines rather than absolute rules when determining if a scoring model is acceptable for operational use. All the performance standards are applied to the independent evaluation sample used to validate the scoring models. Following is a list of the evaluation criteria, along with a description for each criterion and associated statistics. If there are multiple statistics and/or evaluations under a single criterion, they are listed as subheadings.

- ***Construct evaluation.*** Automated scoring capabilities, in general, are designed with certain assumptions and limitations regarding the tasks they will score. It is critical to review the goals of the assessment as supported by the

automated scoring capability. Therefore, the initial step in any prospective use of automated scoring is the evaluation of fit between the goals and design of the assessment (or other use of automated scoring) and the design of the capability itself. The process includes a comparison of the construct of interest with that represented by the capability via a systematic review of the task design specifications, scoring rubrics with selected exemplars, human scoring rules, score reporting goals, and claims and disclosures about assessment scores. This evaluation is qualitative and, hence, no empirical performance standards and/or statistics are associated with this evaluation criterion.

- ***Association with human scores.*** Automated scoring capability has to meet the performance standards for operational use. The performance is evaluated against the human scores using various statistics. Absolute agreement of automated scores with human scores has been a typical measure of the quality of automated scoring. Although it is common to report absolute agreements as percentages of cases with exact agreements and exact-plus-adjacent agreements, in evaluation of e-rater for an assessment, these criteria are scale dependent (i.e., values will be expected to be higher by chance on a 4-point scale than on a 6-point scale) and sensitive to base distributions (i.e., tendencies of human scores to use some score points much more frequently than others). Therefore, they are reported in combination with other more robust statistics, such as *quadratic weighted kappa* (QWK) and Pearson correlations.

  Specifically, the preferred QWK value for *acceptable* consistency of automated and human scores currently is 0.70 (rounded normally). This value was derived on the conceptual basis that it represents the *tipping point* at which signal outweighs noise in the prediction. The identical standard of 0.70 has been adopted for Pearson correlations. It should be noted that the results from QWK and Pearson correlations are not identical, as kappa is computed on the basis of values of e-rater that normally are rounded to the nearest scale score point, while the correlation is computed on the basis of unrounded values (e-rater scores are provided unrounded so that when multiple prompts are combined for a reported score, the precise values can be combined and rounded at the point of scaling rather than rounding prior to summation). It is worthwhile to note that since e-rater is calibrated to empirically optimize the prediction of human scores, the expected performance of e-rater against this criterion is bounded by the performance of human scoring. That is, if the interrater agreement of independent human raters is low, especially below the 0.70 threshold, then automated scoring is disadvantaged in demonstrating this level of performance, not because of any particular failing of automated scoring, but because of the inherent unreliability of the human scoring upon which it is both modeled and evaluated. Therefore, the interrater agreement among human raters is commonly evaluated as a precursor to automated scoring modeling and evaluation. Consequently, additional measures for the quality of automated scores relative to the quality of the human scores are included in the evaluation framework as well. Two such measures are described next.

  Degradation. Another measure of performance in relationship with human scores is degradation, which recognizes the inherent relationship between the reliability of human scoring and the performance of automated scoring. The agreement between human and automated scores is compared to the agreement between two human scores, and it is a concern when the automated-human scoring agreement is more than 0.10 *lower*, in either QWK or correlation, than the human–human agreement.

  This performance standard prevents circumstances in which automated scoring may reach the 0.70 threshold but still be notably deficient in comparison with human scoring. It should be noted that, in practice, occasionally cases are observed in which the automated-human agreement for a particular prompt has been slightly less than the 0.70 performance threshold but very close to a borderline performance for human scoring (e.g., an automated-human weighted kappa of 0.68 and a human–human kappa of 0.71), and such models have been approved for operational use on the basis of being highly similar to human scoring and consistent with the purpose of the assessment for which they are used. Similarly, it is common to observe automated-human absolute agreements that are *higher* than the human–human agreements for prompts that primarily target general writing quality.

  Standardized mean score difference. A third criterion for association of automated scores with human scores is that the standardized mean score difference (standardized using the standard deviation of the distribution of human scores) between the human scores and the automated scores cannot exceed 0.15. This standard ensures that the means of the automated and human score distributions are comparable.

- ***Association with external variables.*** There are problems and concerns with human scoring that represent a range of potential pitfalls, including halo effects, fatigue, tendency to overlook details, and problems with consistency of scoring across time (Braun, 1988; Daly & Dickson-Markman, 1982; Hales & Tokar, 1975; Hughes & Keeling,

1984; Hughes, Keeling, & Tuck, 1980a, 1980b, 1983; Lunz, Wright, & Linacre, 1990; Spear, 1997; Stalnaker, 1936). Therefore, it is of relevance to investigate more than just the consistency with human scores and also to evaluate the patterns of relationship of automated scores, compared to their human counterparts, with scores from external measures. Scores on other test sections to examine within-test relationships, and external scores, such as scores from self-reported measures of interest (e.g., grades in English class, academic majors), are some examples that are used for this purpose. It should be noted that the external scores that are available are typically not a direct external measure of exactly the same construct, which often poses some challenges for interpretation.

- ***Subgroup differences.*** In evaluating fairness of automated scoring, the goal is to investigate any unwarranted impact of substituting a human rater with an automated score on subgroups of examinees. Due to lack of a suitable differential item functioning measure for this purpose, two approaches have been proposed and implemented as measures of fairness for e-rater, specifically, to identify patterns of systematic differences in the distribution of scores between human and automated scoring for subgroups at the reported score level. The first is extending the flagging criterion of standardized mean score differences from the prompt-level analysis discussed above to the evaluation of subgroup differences. A more stringent standard of 0.10 has been adopted at ETS and is applied to all subgroups of interest.

  The second approach is examination of differences in the predictive ability of automated scoring by subgroup. This consists of two classes of prediction that are likewise related to the standards and processes discussed above. First is to compare an initial human score and the automated score in their ability to predict the score of a second human rater by subgroup. The second type of prediction is comparing the automated and human score ability to predict an external variable of interest by subgroup. The approaches and the methods can be selected and applied as appropriate for programs and assessments.
- ***Operational impact analysis.*** The final step is to determine the impact of introducing automated scoring on the aggregate reported score level for the writing section. This is evaluated by simulating the score for the evaluation sample that would result from substituting an automated score for a human score and by determining the distribution of changes in reported scores that would result from such a substitution. This lends an additional opportunity to predict the performance of scoring under the proposed model (automated and human) by comparing it to that of the traditional model (two human raters). The types of empirical analyses conducted for this purpose include an examination of the rate and degree of raw and scaled score differences (as well as classification rates if applicable) resulting from the change, and the degree of differences in association of reported scores to other test scores and external criteria, both at the overall and at the subgroups level. Additionally, the number of second and third human readings is compared under the human-automated and human–human models as a measure of efficiency. All the analyses in this stage allow for the consideration of issues in scale continuity and other factors that may bear on the decision to implement automated scoring.

Regarding adjudication thresholds, alternative thresholds are considered for the definition of *discrepancy* when evaluating the operational agreement between automated and human scores. In human scoring, it is common practice for most scoring scales in high-stakes programs that use double-human scoring to consider scores that are one point apart (e.g., one rater issuing a 3 and the other a 4) to be in agreement under the interpretation that reasonable judges following the rubric may differ, especially when evaluating a borderline submission. Typically, when two human scores are considered discrepant, an adjudication process occurs in which additional human raters are used, and a resolution process is followed to determine the final reported score. These adjudication and resolution processes vary substantially by program and are sometimes conditional on the particular distribution of initial human scores produced. In the implementation of automated scoring with precise values recorded (decimal values), a wider range of options is available for defining agreement, each of which has implications for the extent to which the results of automated scoring influence the final reported scores and, therefore, the ultimate evaluation of impact under the procedures defined above. For example, a 0.5 agreement threshold between human and automated score is more conservative than a 1.5 threshold and will likely result in fewer cases receiving automated scores. The final decision on the acceptable threshold is made by the test program.

Data

More than 68,000 operational responses across 35 items were drawn from the available test records from July 2009 to June 2010, which resulted in roughly 2,000 responses per essay item. Along with the two human rater scores for each essay,

several additional variables were included for analysis, such as test-taker background variables (gender, ethnicity, English as their first language) and other Praxis I test scores on the math and reading sections. The data provided by the Praxis program met all the standards for automated scoring model building and evaluation as described previously.

Methods

The e-rater v10.1 was used for the study. This version of e-rater had 11 features, nine for measuring aspects of general writing quality, including the positive feature on correct use of collocations and prepositions newly introduced in this version, and the two content prompt-specific vocabulary use features. The sentence variety feature was introduced as a new macrofeature in e-rater v13.1 and was, therefore, not included in version 10.1. However, it should be noted that during the annual engine upgrade process each year, new models are built and evaluated using the latest e-rater version for all high- and low-stakes assessments using e-rater for operational scoring. The initial evaluations under v10.1 were, therefore, followed by an upgrade of the scoring model under e-rater v13.1, the current version at the time of writing this report, as well as refresh of the data (used for model building) by addition of more recent data for a small subset of the items. No new prompts were added to the pool of 33 prompts from 2010; however, based on availability, data were refreshed for 9 of the 33 prompts by adding 400 cases for each of these prompts (200 each to model build and cross-validation sets). The results for the updated model are discussed in the report after presenting all the results from initial evaluations.

The review of task design, scoring rubric, human scoring rules, reporting goals, and claims and disclosures for the assessment were made in conjunction with the Praxis program as the study progressed. The expert test developers for the Praxis program evaluated the construct measured by the scoring rubric for the Praxis I essay tasks against the construct represented by e-rater features to determine overlap and/or consistency. The scoring rubric is included in Table A1.

Prompt-specific and generic scoring models were built and evaluated for the Praxis I essay response data from 2009 to 2010 using e-rater v 10.1. The sample size was 200 test takers per prompt for the model-build set for all model types, and the remaining number of responses for each prompt determined the sample size for the evaluation set. The sample size for the evaluation set for each prompt can be found in the appendix tables reporting results for each model at the prompt level.

Agreement statistics for automated scores with human scores were computed for all e-rater models built and evaluated for the Praxis essay response data. For computing exact and adjacent agreement percentages and QWK statistics, the raw e-rater scores are first brought into range to align with the score scale (1–6 for Praxis I essay tasks) and rounded to integers for comparison against the integer human scores. For other agreement statistics, such as Pearson correlation and standardized mean score differences, unrounded e-rater scores aligned with the score scale are used for comparison with human scores. A summary of the flagging criteria and conditions for evaluating model performance, explained previously under the evaluation criteria, is included in Table A2. The proposed model for implementation was then subjected to the remaining evaluation criteria of association with external variables, subgroup differences, agreement thresholds for adjudication, and operational impact analysis.

Various thresholds for allowable discrepancy levels between e-rater and human scores are examined to maximize cost savings related to the use of a second human grader while ensuring valid e-rater scores with acceptable agreement levels with human scores, comparable correlations with external measures, and minimal subgroup differences. The allowable discrepancy threshold between the two human scores on the Praxis I essay task is 1 point. Scores discrepant by more than 1 point (that is, apart by 2 or more points as outlined previously under Praxis I scoring rules) are routed to a third human rater. Since e-rater produces real values unlike human scores, which are restricted to integer values, scores greater than 1 but less than or equal to 1.4999 are rounded down to 1 under normal rounding rules. Hence, replicating the Praxis I scoring rules, a contributory model at threshold of 1.5 was chosen for evaluating the impact of including e-rater in operational scoring for the Praxis I essay task.

The impact of implementation on operational or reported scores was determined by examining the change in scale score distribution and pass-fail rates at the overall and subgroup level, upon substituting one human score with an e-rater score and computing the writing score (referred to as the simulated score) under the preferred implementation model (contributory score with 1.5 agreement threshold). A 95% confidence interval was constructed around the mean reported scale score for each subgroup. Further, as noted in the introduction, every state or agency that uses Praxis I test scores for licensing decisions sets its own requirements and passing score. Therefore, pass-fail indicators were created for all test takers based on the designated institute state recipient (state/institute) and the state's certification procedure, and change in pass-fail rates as a result of substituting one human score with an e-rater score was examined at the state level.

The anticipated number of second human ratings for scores based on all-human scoring and scores based on human and e-rater combined were compared. For two human scores, the third rating was provided by a third human rater when the human scores differ by 2 or more points. For one human and one e-rater score, the third rating was provided by a second human rater when the human and e-rater scores differ by 1.5 points or more, or if e-rater issued an advisory flag.

Results

Construct Relevance

The test developers for the Praxis program and the writing experts at ETS evaluated the construct measured by the scoring rubric for the Praxis I essay tasks against the construct represented by e-rater features and determined a sufficient overlap and/or consistency. The Praxis I essay tasks require test takers to clearly and effectively state a thesis, develop and organize ideas with reasons and examples, use effective sentence construction and variety, and display facility with good language use by producing writing that is free of grammatical and mechanical errors. The organization and development features in e-rater effectively evaluate the structural and transition elements, such as thesis, as emphasized by the rubric. The use of language facility (grammar, usage, and mechanics) and sentence variety are evaluated by e-rater by means of a variety of microfeatures that measure sentence-level errors (e.g., run-on sentences and fragments), grammatical errors (e.g., subject-verb agreement), and types of clauses and phrases. Based on this mapping, it was concluded that all the key elements of the scoring guide are sufficiently measured by e-rater features. The similarity between essay tasks and the scoring guides across Praxis I, GRE, and TOEFL further lent supporting evidence for sufficient construct coverage under e-rater for scoring Praxis I writing tasks.

Advisory Analyses

All advisories were evaluated against human 1 (H1) ratings, individually and sequentially, and, as a result, four flags were identified for use in operational scoring—reuse of language (flag 2), less development of the key concepts than other essays written on the topic (flag 4), unidentifiable organizational elements (flag 256), and excessive number of problems (large number of grammar, usage, and mechanics errors; flag 512)—along with filters to remove responses that were too short (less than 25 words or two sentences; flag 64) and/or too long (greater than 1,000 words; flag 128). The use of the recommended advisory rules resulted in a very small number of additional cases, 3%, requiring double-human scoring. Table 2 reports the flagging rates conditional on the first human score.

Agreement With Human Scores

Evaluation of the differences in raw scores under human–e-rater (H1-e-rater) scoring compared to human–human (H1–H2) scoring was conducted. Recall that the raw e-rater scores are produced on a continuous scale under the linear regression model. Table 3 shows results for QWK, Pearson correlations, standardized mean score differences, and degradation of e-rater-human agreement from human–human agreement.

For the operational Praxis I essay tasks, there was a 68% exact agreement and 99% exact plus adjacent agreement between scores given by two human raters. The QWK and correlation for scores by human raters was 0.74.

Table 2 Flagging Rates for the Praxis I Essay Items

Human score	No flag	Flag	% Flagged[a]	Row sum
1	13	7	35	20
2	389	32	8	421
3	2,445	95	4	2,540
4	3,062	75	2	3,137
5	794	4	1	798
6	84	0	0	84
Total	6,787	213	3	7,000

[a]Decimal values rounded up.

Table 3 Agreement With Human Scores

		H1 by H2										H1 by e-rater (rounded to integers)										H1 by e-rater (unrounded)						Degradation	
		H1		H2		Stats						H1		e-rater		Stats						H1		e-rater		Stats		QWK	*r*
	N	*M*	*SD*	*M*	*SD*	*SMD*	*K*	*QWK*	% agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	*QWK*	% agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*r*	h1er_rnd – h1h2	h1er – h1h2
Generic	737	3.68	0.81	3.67	0.81	−0	0.51	0.74	68	99	0.74	3.68	0.81	3.68	0.86	−0	0.56	0.78	71	99	0.78	3.68	0.81	3.67	0.8	−0.01	0.82	0.04	0.08
Prompt specific	460	3.68	0.82	3.67	0.81	−0	0.51	0.74	68	99	0.74	3.68	0.82	3.68	0.86	0	0.55	0.78	70	99	0.78	3.68	0.82	3.67	0.81	−0.01	0.82	0.04	0.08

Note. adj = adjacent; H1 = Human 1; H2 = Human 2; h1er_rnd = H1 by e-rater (rounded to integers); h1er = H1 by e-rater (unrounded); *K* = kappa; *M* = mean; *QWK* = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; *SMD* = standardized mean difference.

Table 4 Human and e-rater Score Correlations With Other Measures

	Praxis I reading scaled score	Praxis I mathematics scaled score	Praxis I MC score	Praxis I writing scaled score
Human score	0.32	0.31	0.36	0.73
e-rater score	0.30	0.32	0.35	0.69
Unsigned difference	0.02	0.01	0.01	0.04

The agreement between human and automated scoring was evaluated under both generic and prompt-specific models. The human–e-rater agreement was comparable to the agreement between human raters under the preferred generic e-rater model. All the standard e-rater evaluation criteria were sufficiently met at the aggregated level with improved QWK and correlations for e-rater with human score (QWK = 0.78, $r = 0.82$). The agreement between e-rater and human scoring met the expectation of minimal degradation (decrease in associate measures of less than .10) from human–human agreement. In fact, the degradation was actually an increase, in that the difference between the interrater agreement of humans (correlation of .74) was lower than the automated-human agreement (correlation of 0.82). It reflected an average improvement in agreement of 0.08, while QWK improved by 0.04 when using an automated rating. The mean score differences between human and e-rater across prompts was −0.01 at the overall level, well within the accepted limit of 0.15 of a standard deviation (using the standard deviation of the human-score distribution).

The tables reporting results for each model at the prompt level are included in Appendix B (Tables B1 and B2). The bold numbers in these tables fail to meet the threshold values for the respective statistic. It should be noted that all the threshold values are evaluated to four decimal places for flagging purposes. For the results at the prompt level, the QWK and correlation values for the two human scores, which serve as the baseline for human with e-rater agreement statistics, failed to meet the acceptable threshold for one of the 35 prompts. For the automated scores, all the performance standards were met under the generic model, except for two of the 35 prompts for which human and e-rater mean score difference exceeded the threshold of 0.15. These two prompts (VB315230 and VB398189) were removed, and the generic model was rebuilt and evaluated on the remaining 33 prompts. The prompt-level results reported in Tables B1 and B2 are for the pool of 33 prompts.

Based on the results for the evaluation criteria at the aggregate and the prompt level, the preferred generic model fared well for Praxis I essay tasks. Because of the unusual performance of prompts VB315230 and VB398189, these two prompts were recommended to be removed from the operational prompt pool with e-rater in use for scoring. The subsequent analyses use e-rater scores produced from the generic scoring model built on the pool of 33 prompts for validity evidence supporting use of automated scoring for Praxis I essay task.

Association With External Measures

We computed correlations for e-rater and human scores with external measures such as scores on other test sections (Praxis I reading and mathematics), the total right score for the MC section of the writing test, and the total scaled score for the writing section. Table 4 reports the association of the rounded e-rater scores from the generic model and human scores with the scores on external measures, that is, scores on other test sections (Praxis I reading and mathematics), the total right score for the MC section of the writing test, and the total scaled score for the writing section.

The correlations with the external variables for e-rater and human scores were comparable, with correlations for e-rater scores lower for reading and MC test scores by 0.01 and for the scaled score for the writing test by 0.04, and correlations for human scores lower for the mathematics test score by 0.01.

Subgroup Differences

Performance of the preferred scoring model was evaluated for systematic differences in e-rater scores across subgroups based on gender, first language, and ethnicity. Table 5 shows the results for QWK, Pearson correlation, standardized mean score difference, and degradation of e-rater-human agreement from human–human agreement for the one subgroup of the Asian Pacific Islander ethnic group, the only subgroup with absolute value of mean score difference greater than the threshold of 0.10 ($SMD = 0.11$, with higher e-rater scores). The standardized mean score differences at the subgroup level

Table 5 Agreement With Human Scores for Asian Pacific Islander Test Takers

		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater		Stats				e-rater		Stats		QWK	r
Subgroup	N	M	SD	M	SD	SMD	K	QWK	% agree	% adj agree	r	M	SD	K	QWK	% agree	% adj agree	M	SD	SMD	r	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Asian Pacific Islander	1,549	3.54	0.88	3.51	0.87	−0.03	0.5	0.75	66.04	98.84	0.75	3.63	0.86	0.52	0.76	67.14	98.52	3.63	0.82	**0.11**	0.8	0	0.05

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1; H2 = Human 2; K = kappa; M = mean; QWK = quadratic-weighted kappa; r = correlation; SD = standard deviation; SMD = standardized mean difference.

Table 6 Total Writing Score Correlations With Other Measures Under Contributory Score Model for e-rater

	Praxis I reading scaled score	Praxis I mathematics scaled score	Praxis I MC score	Praxis I writing scaled score
Essay score (all human)	0.34	0.33	0.39	0.78
Essay score (with e-rater)	0.33	0.33	0.38	0.76
Unsigned difference	0.01	0	0.01	0.02
Raw weighted composite writing score (all human)	0.63	0.56	0.87	1.00
Raw weighted composite writing score (with e-rater)	0.63	0.56	0.87	0.98
Unsigned difference	0	0	0	0.02

were evaluated again by simulating the writing score (using e-rater under the preferred model for implementation) to assess the impact on reported scores.

It should be noted that for subgroups with small sample sizes (less than 1,000 test takers), any differences around the value of 0.10 are not considered for further formal review. Results for other subgroups that were evaluated (gender, English as first language, and other ethnic groups) are included in Appendix B (Table B3). Human–e-rater agreement was slightly better than human–human agreement for the African American and American Indian group of test takers.

Models for Implementation

Table 6 reports the correlations of final scores for the Praxis I writing test simulated under the contributory score model at 1.5 threshold with other measures. The simulation entails substituting one human score with the e-rater score, and computing the essay score and the total writing score under the contributory model with 1.5 agreement threshold between human and e-rater score. Compared to the operational writing score produced using two or more human ratings, the new simulated writing scores produced using combined e-rater and human scores show fairly similar association with scores on other Praxis I test sections, the total right score for MC section, and the total scaled score for writing. As mentioned previously, the standardized mean differences at the subgroup level were examined again, and there were no subgroup differences of concern under this model (Table B4).

Impact of Implementation

To examine the impact of implementing e-rater, the score distribution for operational reported scale scores based on human scores and the simulated scores using e-rater were compared. The rate of perfect agreement at the overall level was approximately 50%, with 1 point discrepancies occurring approximately 43% of the time. These discrepancies are evenly divided between increases and decreases in scale scores, with 25% increasing by 1–2 points and 24% decreasing by 1–2 points (see Table B5). Also, at the reported score level, there was no demographic subgroup for which mean simulated score differed from the operational mean by more than chance variation. All simulated mean scores fell within the 95% confidence interval constructed around the mean reported scale score for each subgroup (see Table B6).

Overall, there was no change in the passing status for almost 96% of test takers. At the subgroup level (gender, ethnicity, and English as first language), there was no change in the passing status of 94–96% of test takers. Based on the student's designated institute recipient (state level), there was no change in the passing status for 92–100% of students. When status did change, candidates were approximately split between changing from pass to fail status or vice versa (see Tables B7 and B8).

On comparing the anticipated number of second human ratings for scores based on all-human scoring and scores based on human and e-rater combined, only 4% of cases were identified as the ones that will need more than one human score when using e-rater, suggesting more efficient use of human raters and reduced score turnaround time.

The results for the upgraded model under the e-rater engine v13.1 are highly similar to the results observed for the original model, while eliminating a previous violation of standardized mean score difference threshold at the subgroup level ($SMD = 0.11$ for Asian Pacific Islander group). And with no new additions to the prompt pool and relatively minor differences in the computation of the total writing score, it was agreed upon by all stakeholders that the original results

generalize to the new test sufficiently therefore not warranting a full reevaluation. The performance results for the updated model under engine v13.1 are included in Appendix B (Tables B9 and B10).

Conclusion

Prompt-specific and generic scoring models were built and evaluated on Praxis I essay response data from July 2009 to June 2010 using e-rater v10.1. These data comprised about 68,000 essay responses written to 35 prompts; eventually, data for two prompts were excluded from the study on account of poor fit for automated scoring. Criteria for evaluation of e-rater scoring models included level of agreement with human scores, degradation in agreement from human scoring, standardized mean score differences between human and automated scoring, and correlations with external variables (such as scores on other Praxis I test sections, total right score for MC writing, and total scaled score for writing). Based on the evaluation criteria, a generic model was recommended for implementation for operational use. Performance of the generic model was further evaluated across different demographic subgroups. Results revealed adequate performance at the subgroup level, with an exception of discrepancy between e-rater and human scores for Asian Pacific Islander test takers (about one 10th of a *SD* higher than the human score).

The use of e-rater score was investigated as a contributory score. Under the contributory score model, e-rater score was checked for agreement with the first human score within an empirically established range, beyond which a second human score was required. The sum of the human and e-rater scores became the final score for the essay, unless a second human rating was required. Various agreement thresholds were considered under the contributory score model to minimize differences across the subgroups. Discrepancy threshold of 1.5 points between the automated and the human score was selected to yield performance as similar as possible to double-human scoring, and with significant savings in second human ratings. The observed mean score difference for the one subgroup was mitigated at the overall reported score level under the chosen implementation model.

As a result of this study, a generic e-rater scoring model producing an automated score as a replacement for a second human score when the agreement threshold was within 1.5 units of the first human score was recommended for operational use in the Praxis I program. Based on the results under the more recent version that meet all performance criteria, the updated model was determined suitable for operational use. As part of future efforts, it will be critical to monitor and evaluate the operational performance of the scoring model from time to time to account for any changes in the examinee characteristics, human scoring trends over time, and new feature developments and enhancements in the e-rater engine.

Acknowledgments

The authors wish to thank Yigal Attali, Brent Bridgeman, Tim Davey, Neil Dorans, Marna Golub-Smith, Shelby Haberman, and Charles Lewis for their assistance in interpretation of the results; Kevin Larkin, Danielle Siwek, Bob Smith, and Adele Tan for advising on psychometric policies and procedures for the program; Nancy Glazer for advising on human scoring policies and procedures for the program; Rick Tanenbaum for advising on standard-setting policies and procedures and cut scores for different states; Kevin Cureton, Linda Tyler, and the *PRAXIS*® program for providing the data and sharing their expert knowledge of the program; and Slava Andreyev, Bob D'Addario, Matthew Duchnowski, Laura Ridolfi, Jonathan Steinberg, Waverely VanWinkle, and Vincent Weng for their assistance with the data and analyses.

References

Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated scoring. *Journal of Technology, Learning, and Assessment, 10*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603/1455

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650/1492

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Bernstein, J., De Jong, J., Pisoni, D., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings in InSTIL2000* (pp. 57–61). Dundee, Scotland: University of Abertay.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, *13*, 1–18.

Breyer, F. J., Attali, Y., Williamson, D. M., Ridolfi, L., Ramineni, C., & Duchnowski, M. (2014). *A study of the use of e-rater for the analytical writing measure of the GRE revised General Test* (Research Report No. RR-14-24). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12022

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Hillsdale, NJ: Erlbaum.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada. Retrieved from www.ets.org/Media/Research/pdf/erater_ncmefinal.pdf

Callear, D., Jerrams-Smith, J., & Soh, V. (2001). CAA of short non-MCQ answers. In *Proceedings of the 5th International CAA Conference* (pp. 55–69). Loughborough, England: Loughborough University.

Chevalier, S. (2007). Speech interaction with Saybot player, a CALL software to help Chinese learners of English. In *Proceedings of The International Speech Communication Association Special Interest Group on Speech and Language Technology in Education (SLaTE)*. Farmington, PA: ISCA SIG.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Culham, R. (2003). *6 + 1 traits of writing: The complete guide*. New York, NY: Scholastic.

Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, *19*, 309–316.

Educational Testing Service. (2014). *Praxis I® overview.* Retrieved from http://www.ets.org/praxis/about/praxisi/praxisi_overview

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., … Cesari, F. (2000). The SRI EduSpeak™ system: Recognition and pronunciation scoring for language learning. In P. Delcloque (Ed.), *Proceedings of InSTILL* (pp. 123–128). Dundee, Scotland: University of Abertay.

Hales, L. W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, *12*, 115–117.

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement*, *21*, 277–281.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980a). Essay marking and the context problem. *Educational Research*, *22*, 147–148.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1980b). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement*, *17*, 131–135.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring. *Educational and Psychological Measurement*, *43*, 1047–1050.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Hillsdale, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331–345.

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In *Proceedings of the Sixth International Computer Assisted Assessment Conference* (pp. 233–249). Loughborough, England: Loughborough University.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *48*, 238–243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, *14*(2), 210–225.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Hillsdale, NJ: Erlbaum.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02158.x

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012a). *Evaluation of e-rater for GRE issue and argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02284.x

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater scoring engine for TOEFL Independent and Integrated prompts* (Research Report No. RR-12-06). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02288.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39.

Risse, T. (2007, September). *Testing and assessing mathematical skills by a script based system.* Paper presented at the 10th International Conference on Interactive Computer Aided Learning, Villach, Austria.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment, 4*(4). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1651/1493

Sargeant, J., Wood, M. M., & Anderson, S. M. (2004). A human-computer collaborative approach to the marking of free text answers. In *Proceedings of the 8th International CAA Conference* (pp. 361–370). Loughborough, England: Loughborough University.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum.

Singley, M. K., & Bennett, R. E. (1998). *Validation and extension of the mathematical expression response type: Applications of schema theory to automatic scoring and item generation in mathematics* (GRE Board Professional Report 93-24P). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1997.tb01740.x

Spear, M. (1997). The influence of contrast effects upon teachers' marks. *Educational Research*, *39*, 229–233.

Stalnaker, J. M. (1936). The problem of the English examination. *Educational Record*, *17*, 41.

Sukkarieh, J. Z., & Pulman, S. G. (2005). Information extraction and machine learning: Auto-marking short free text responses to science questions. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)* (pp. 629–637). Amsterdam, The Netherlands: AIED.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, *31*(1), 2–13.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1016/j.csl.2010.06.001

Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL* (pp. 216–223). New York, NY: The Association for Computational Linguistics.

Appendix A. Organization and Flagging Criteria of e-rater, Scoring guide

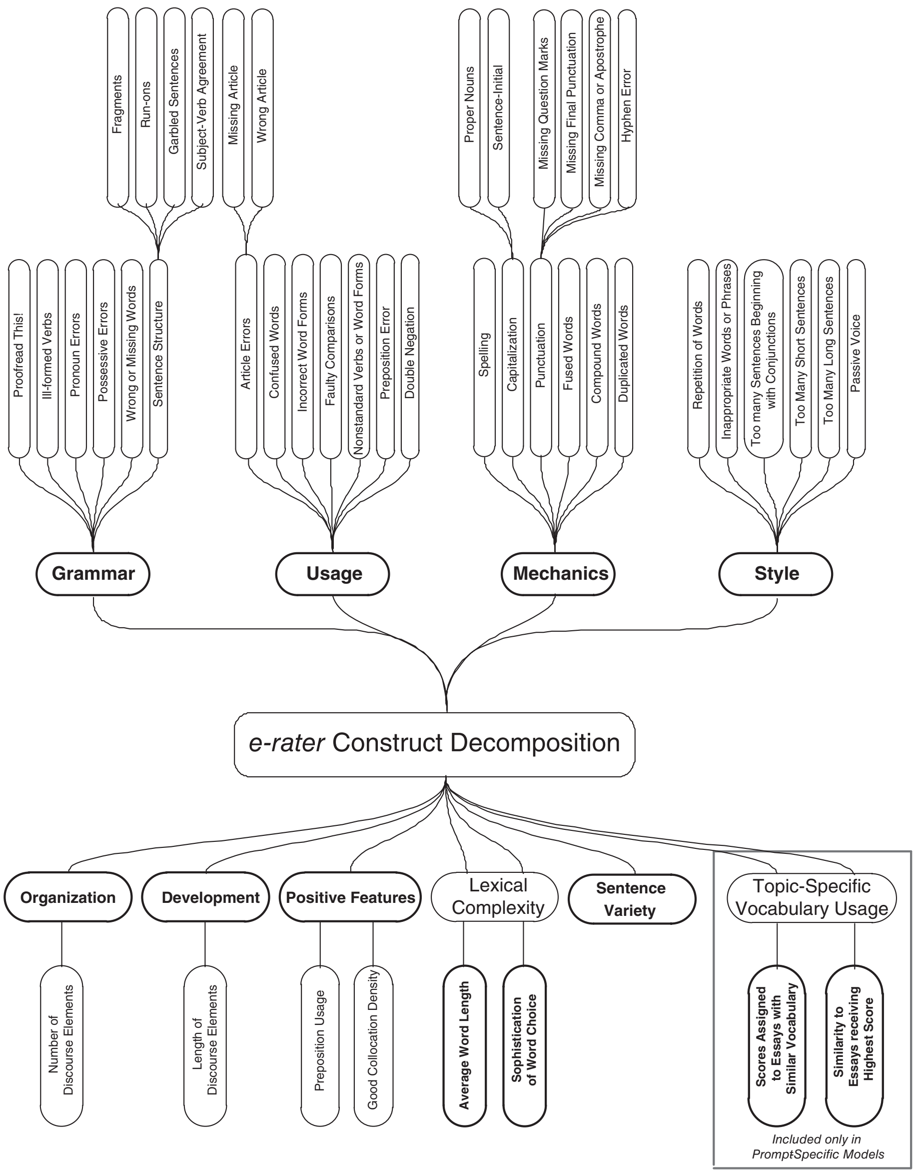

Figure A1 Organization and construct coverage of e-rater v13.1 Adapted from "Evaluating the Construct Coverage of the e-rater Scoring Engine" (Research Report No. RR-09-01), by T. Quinlan, D. Higgins, and S. Wolff, 2009, p. 9. Copyright 2009 by Educational Testing Service, Princeton, NJ.

Table A1 Praxis I Essay Scoring Guide

Score	Description
6	*A 6 essay demonstrates a high degree of competence in response to the assignment but may have a few minor errors.* An essay in this category: ● states or clearly implies the writer's position or thesis ● organizes and develops ideas logically, making insightful connections between them ● clearly explains key ideas, supporting them with well chosen reasons, examples, or details ● displays effective sentence variety ● clearly displays facility in the use of language ● is generally free from errors in grammar, usage, and mechanics

Table A1 (continued)

Score	Description
5	*A 5 essay demonstrates clear competence in response to the assignment but may have minor errors.* An essay in this category: ● states or clearly implies the writer's position or thesis ● organizes and develops ideas clearly, making connections between them ● explains key ideas, supporting them with relevant reasons, examples, or details ● displays some sentence variety ● displays facility in the use of language ● is generally free from errors in grammar, usage, and mechanics
4	*A 4 essay demonstrates competence in response to the assignment.* An essay in this category: ● states or implies the writer's position or thesis ● shows control in the organization and development of ideas ● explains some key ideas, supporting them with adequate reasons, examples, or details ● displays adequate use of language ● shows control of grammar, usage, and mechanics, but may display errors
3	*A 3 essay demonstrates some competence in response to the assignment but is obviously flawed.* An essay in this category reveals one or more of the following weaknesses: ● limited in stating or implying a position or thesis ● limited control in the organization and development of ideas ● inadequate reasons, examples, or details to explain key ideas ● an accumulation of errors in the use of language ● an accumulation of errors in grammar, usage, and mechanics
2	*A 2 essay is seriously flawed.* An essay in this category reveals one or more of the following weaknesses: ● no clear position or thesis ● weak organization or very little development ● few or no relevant reasons, examples, or details ● frequent serious errors in the use of language ● frequent serious errors in grammar, usage, and mechanics
1	*A 1 essay demonstrates fundamental deficiencies in writing skills.* An essay in this category: ● contains serious and persistent writing errors or ● is incoherent or ● is undeveloped

Table A2 Flagging Criterion and Conditions

Flagging criterion	Flagging condition
Quadratic-weighted kappa between e-rater score and human score	Quadratic-weighted kappa less than 0.70
Pearson correlation between e-rater score and human score	Correlation less than 0.70
Standardized difference between e-rater score and human score	Standardized difference greater than 0.15 in absolute value
Notable reduction in quadratic-weighted kappa or correlation from human–human to automated-human	Decline in quadratic-weighted kappa or correlation of greater than 0.10
Standardized difference between e-rater score and human score within a subgroup of concern	Standardized difference greater than 0.10 in absolute value

Note. All the threshold values are evaluated to 4 decimal values for flagging.

Appendix B. Agreement statistics and Operational impact

Table B1 Agreement With Human Scores at Prompt Level: Generic Model (33 Prompts)

		H1 by H2										H1 by e-rater (rounded to integers)										H1 by e-rater (unrounded)						Degradation	
		H1		H2		Stats						H1		e-rater		Stats						H1		e-rater		Stats		QWK	r
Prompt	*N*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*r*	h1er_rnd – h1h2	h1er – h1h2
VB315223	750	3.67	0.82	3.67	0.8	0	0.47	0.71	65	99	0.72	3.67	0.82	3.67	0.84	0	0.59	0.79	73	99	0.79	3.67	0.82	3.66	0.79	−0.01	0.82	0.08	0.11
VB315224	737	3.73	0.81	3.75	0.83	0.02	0.54	0.76	70	99	0.76	3.73	0.81	3.66	0.88	−0.08	0.56	0.79	71	100	0.79	3.73	0.81	3.66	0.85	−0.08	0.84	0.03	0.08
VB315225	737	3.76	0.79	3.75	0.79	−0.02	0.54	0.73	70	99	0.73	3.76	0.79	3.74	0.84	−0.02	0.57	0.78	72	100	0.78	3.76	0.79	3.74	0.78	−0.03	0.82	0.06	0.1
VB315226	718	3.73	0.84	3.75	0.84	0.02	0.5	0.75	67	99	0.75	3.73	0.84	3.77	0.84	0.05	0.6	0.8	74	100	0.81	3.73	0.84	3.77	0.8	0.05	0.83	0.06	0.08
VB315227	738	3.72	0.79	3.75	0.8	0.03	0.53	0.73	70	99	0.73	3.72	0.79	3.8	0.85	0.1	0.56	0.75	71	99	0.76	3.72	0.79	3.82	0.79	0.12	0.79	0.03	0.07
VB315228	745	3.6	0.78	3.58	0.79	−0.03	0.51	0.72	69	99	0.72	3.6	0.78	3.58	0.86	−0.03	0.57	0.78	72	99	0.78	3.6	0.78	3.58	0.8	−0.03	0.8	0.06	0.08
VB315231	740	3.78	0.81	3.82	0.81	0.05	0.52	0.74	69	99	0.75	3.78	0.81	3.83	0.84	0.06	0.55	0.76	70	99	0.76	3.78	0.81	3.82	0.77	0.06	0.8	0.02	0.05
VB315232	731	3.53	0.78	3.52	0.8	−0.01	0.48	0.7	67	99	0.7	3.53	0.78	3.43	0.85	−0.12	0.48	0.74	66	100	0.74	3.53	0.78	3.42	0.8	−0.13	0.78	0.03	0.08
VB315233	724	3.67	0.77	3.71	0.78	0.06	0.48	0.71	67	99	0.71	3.67	0.77	3.66	0.81	0	0.55	0.75	71	99	0.75	3.67	0.77	3.65	0.76	−0.02	0.81	0.05	0.1
VB315234	731	3.7	0.81	3.7	0.79	−0.01	0.47	0.72	66	99	0.72	3.7	0.81	3.72	0.85	0.01	0.53	0.77	69	99	0.77	3.7	0.81	3.71	0.8	0.01	0.8	0.05	0.08
VB315235	741	3.54	0.82	3.54	0.83	0	0.5	0.74	67	99	0.74	3.54	0.82	3.44	0.9	−0.13	0.54	0.79	69	100	0.8	3.54	0.82	3.44	0.85	−0.13	0.83	0.05	0.09
VB384839	734	3.69	0.79	3.7	0.8	0.01	0.5	0.73	68	99	0.73	3.69	0.79	3.73	0.81	0.05	0.5	0.74	68	99	0.74	3.69	0.79	3.74	0.75	0.07	0.8	0	0.06
VB384840	740	3.62	0.81	3.61	0.81	−0.02	0.54	0.75	71	99	0.75	3.62	0.81	3.61	0.87	−0.02	0.55	0.77	70	99	0.78	3.62	0.81	3.6	0.81	−0.03	0.83	0.02	0.08
VB384841	725	3.55	0.83	3.56	0.8	0	0.54	0.77	70	100	0.77	3.55	0.83	3.52	0.83	−0.04	0.56	0.78	71	99	0.78	3.55	0.83	3.5	0.81	−0.06	0.82	0.01	0.05
VB396482	746	3.77	0.83	3.76	0.81	−0.01	0.51	0.72	68	98	0.72	3.77	0.83	3.79	0.86	0.02	0.55	0.77	70	99	0.77	3.77	0.83	3.76	0.81	−0.01	0.81	0.05	0.09
VB396483	754	3.64	0.75	3.65	0.76	0.01	0.51	0.72	69	100	0.72	3.64	0.75	3.64	0.85	0	0.56	0.77	71	100	0.78	3.64	0.75	3.63	0.79	−0.01	0.81	0.05	0.09
VB396484	732	3.57	0.83	3.54	0.81	−0.04	0.54	0.76	70	99	0.76	3.57	0.83	3.55	0.88	−0.02	0.56	0.79	71	100	0.8	3.57	0.83	3.54	0.83	−0.03	0.83	0.03	0.07
VB396485	741	3.61	0.91	3.65	0.86	0.04	0.45	0.72	63	98	0.72	3.61	0.91	3.69	0.89	0.09	0.52	0.78	67	99	0.78	3.61	0.91	3.68	0.85	0.08	0.82	0.06	0.1
VB396486	744	3.76	0.82	3.74	0.8	−0.02	0.5	0.75	68	100	0.75	3.76	0.82	3.76	0.88	0	0.55	0.79	70	100	0.79	3.76	0.82	3.76	0.81	0	0.83	0.04	0.08
VB396487	742	3.62	0.76	3.65	0.77	0.03	0.54	0.74	71	100	0.74	3.62	0.76	3.6	0.86	−0.02	0.59	0.79	74	100	0.8	3.62	0.76	3.6	0.79	−0.04	0.84	0.05	0.09
VB396488	748	3.63	0.8	3.58	0.81	−0.06	0.53	0.74	69	99	0.74	3.63	0.8	3.56	0.86	−0.09	0.58	0.79	72	99	0.79	3.63	0.8	3.56	0.82	−0.09	0.83	0.04	0.09
VB396489	736	3.71	0.84	3.7	0.83	−0.02	0.57	0.78	72	99	0.78	3.71	0.84	3.71	0.87	0	0.59	0.79	72	99	0.79	3.71	0.84	3.71	0.82	−0.01	0.83	0.01	0.05
VB396490	746	3.62	0.84	3.58	0.84	−0.05	0.48	0.74	66	99	0.74	3.62	0.84	3.54	0.9	−0.08	0.5	0.76	66	99	0.76	3.62	0.84	3.55	0.84	−0.08	0.81	0.02	0.07
VB398183	736	3.62	0.84	3.56	0.79	−0.07	0.5	**0.7**	67	98	**0.7**	3.62	0.84	3.57	0.84	−0.06	0.51	0.74	68	99	0.75	3.62	0.84	3.56	0.8	−0.07	0.8	0.05	0.11
VB398184	724	3.69	0.82	3.68	0.85	−0.01	0.53	0.76	69	99	0.76	3.69	0.82	3.74	0.9	0.06	0.59	0.8	72	100	0.81	3.69	0.82	3.75	0.83	0.07	0.84	0.05	0.08
VB398185	728	3.72	0.81	3.65	0.75	−0.09	0.53	0.73	70	99	0.74	3.72	0.81	3.75	0.83	0.03	0.62	0.81	75	100	0.81	3.72	0.81	3.74	0.78	0.03	0.83	0.08	0.1
VB398186	735	3.69	0.82	3.76	0.84	0.09	0.5	0.73	67	99	0.74	3.69	0.82	3.7	0.88	0.02	0.56	0.79	71	99	0.79	3.69	0.82	3.7	0.81	0.01	0.83	0.05	0.09
VB398187	738	3.77	0.8	3.74	0.8	−0.04	0.51	0.74	68	99	0.74	3.77	0.8	3.76	0.85	−0.02	0.58	0.79	73	100	0.79	3.77	0.8	3.75	0.79	−0.02	0.83	0.05	0.09
VB398188	729	3.77	0.8	3.78	0.82	0.02	0.48	0.73	66	100	0.73	3.77	0.8	3.79	0.86	0.03	0.55	0.78	70	100	0.78	3.77	0.8	3.79	0.81	0.04	0.82	0.04	0.09
VB398190	744	3.62	0.81	3.58	0.81	−0.05	0.48	0.73	67	99	0.73	3.62	0.81	3.61	0.84	−0.01	0.54	0.76	70	99	0.76	3.62	0.81	3.61	0.8	−0.02	0.81	0.03	0.08
VB398192	740	3.8	0.84	3.79	0.88	−0.02	0.49	0.75	66	99	0.75	3.8	0.84	3.78	0.85	−0.03	0.56	0.77	71	99	0.77	3.8	0.84	3.77	0.8	−0.04	0.81	0.02	0.06
VB398193	734	3.75	0.84	3.71	0.82	−0.04	0.5	0.73	67	99	0.73	3.75	0.84	3.78	0.85	0.04	0.6	0.8	73	99	0.8	3.75	0.84	3.78	0.81	0.04	0.84	0.08	0.11
VB398194	732	3.82	0.85	3.77	0.81	−0.06	0.47	0.72	66	99	0.72	3.82	0.85	3.84	0.85	0.03	0.55	0.78	70	99	0.78	3.82	0.85	3.84	0.8	0.03	0.82	0.06	0.1

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1; H2 = Human 2; h1er_rnd = H1 by e-rater (rounded to integers); h1er = H1 by e-rater (unrounded); *K* = kappa; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; *SMD* = standardized mean difference.

Table B2 Agreement With Human Scores at Prompt Level: PS Model (33 Prompts)

		H1 by H2										H1 by e-rater (rounded to integers)										H1 by e-rater (unrounded)						Degradation	
		H1		H2		Stats						H1		e-rater		Stats						H1		e-rater		Stats		QWK	r
Prompt	*N*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% Agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*M*	*SD*	*SMD*	*r*	h1er_rnd – h1h2	h1er – h1h2
VB315223	468	3.69	0.84	3.66	0.81	−0.04	0.48	0.72	66	99	0.72	3.69	0.84	3.64	0.78	−0.06	0.56	0.77	71	100	0.77	3.69	0.84	3.63	0.76	−0.08	0.82	0.05	0.1
VB315224	460	3.74	0.79	3.76	0.83	0.03	0.52	0.74	69	99	0.74	3.74	0.79	3.77	0.83	0.04	0.59	0.78	73	99	0.78	3.74	0.79	3.75	0.8	0.02	0.82	0.04	0.07
VB315225	458	3.76	0.8	3.72	0.78	−0.04	0.56	0.75	72	99	0.75	3.76	0.8	3.72	0.87	−0.04	0.54	0.77	69	100	0.78	3.76	0.8	3.73	0.8	−0.03	0.82	0.03	0.07
VB315226	445	3.74	0.83	3.76	0.86	0.02	0.54	0.76	70	99	0.76	3.74	0.83	3.7	0.9	−0.05	0.57	0.8	71	100	0.81	3.74	0.83	3.71	0.86	−0.04	0.84	0.04	0.07
VB315227	460	3.68	0.79	3.73	0.82	0.06	0.5	0.71	68	98	0.72	3.68	0.79	3.71	0.88	0.04	0.58	0.78	72	99	0.78	3.68	0.79	3.71	0.83	0.03	0.81	0.06	0.09
VB315228	467	3.58	0.77	3.58	0.77	0	0.46	**0.69**	66	99	**0.69**	3.58	0.77	3.58	0.87	0.01	0.55	0.76	70	99	0.76	3.58	0.77	3.58	0.78	0.01	0.8	0.06	0.1
VB315231	462	3.74	0.79	3.76	0.78	0.03	0.51	0.74	69	100	0.74	3.74	0.79	3.8	0.83	0.07	0.55	0.77	70	100	0.77	3.74	0.79	3.78	0.77	0.05	0.8	0.03	0.07
VB315232	462	3.55	0.78	3.55	0.8	0	0.5	0.71	68	99	0.71	3.55	0.78	3.6	0.85	0.07	0.5	0.74	67	99	0.74	3.55	0.78	3.6	0.79	0.06	0.78	0.03	0.07
VB315233	451	3.67	0.79	3.73	0.79	0.07	0.49	0.73	68	99	0.73	3.67	0.79	3.7	0.79	0.04	0.54	0.76	71	100	0.76	3.67	0.79	3.68	0.76	0.01	0.81	0.03	0.08
VB315234	453	3.72	0.81	3.72	0.79	0	0.49	0.73	67	100	0.73	3.72	0.81	3.71	0.87	−0.01	0.5	0.76	67	100	0.76	3.72	0.81	3.7	0.83	−0.01	0.81	0.03	0.08
VB315235	463	3.59	0.81	3.55	0.85	−0.05	0.52	0.75	68	99	0.75	3.59	0.81	3.62	0.82	0.03	0.59	0.8	73	100	0.8	3.59	0.81	3.59	0.8	−0.01	0.83	0.05	0.08
VB384839	454	3.65	0.77	3.72	0.78	0.09	0.5	0.72	69	99	0.72	3.65	0.77	3.71	0.9	0.07	0.48	0.74	65	99	0.75	3.65	0.77	3.71	0.85	0.07	0.79	0.02	0.07
VB384840	462	3.64	0.82	3.63	0.8	−0.01	0.53	0.74	69	99	0.74	3.64	0.82	3.65	0.88	0.01	0.54	0.78	69	99	0.78	3.64	0.82	3.63	0.84	−0.02	0.81	0.04	0.08
VB384841	454	3.57	0.84	3.57	0.8	0	0.51	0.76	68	100	0.76	3.57	0.84	3.58	0.84	0.01	0.57	0.78	71	99	0.78	3.57	0.84	3.59	0.82	0.02	0.81	0.02	0.05
VB396482	467	3.81	0.79	3.84	0.77	0.04	0.51	0.71	69	99	0.71	3.81	0.79	3.88	0.83	0.08	0.53	0.75	69	99	0.75	3.81	0.79	3.85	0.78	0.05	0.8	0.04	0.09
VB396483	471	3.66	0.75	3.66	0.76	0.01	0.51	0.73	70	100	0.73	3.66	0.75	3.66	0.84	0.01	0.54	0.76	70	100	0.77	3.66	0.75	3.66	0.77	0	0.81	0.03	0.08
VB396484	465	3.55	0.83	3.54	0.82	−0.02	0.54	0.76	70	99	0.76	3.55	0.83	3.56	0.85	0.01	0.56	0.79	71	100	0.79	3.55	0.83	3.56	0.8	0.02	0.83	0.03	0.07
VB396485	459	3.59	0.91	3.61	0.86	0.02	0.44	0.72	62	98	0.72	3.59	0.91	3.62	0.97	0.03	0.47	0.78	63	100	0.78	3.59	0.91	3.6	0.92	0.01	0.82	0.06	0.1
VB396486	468	3.76	0.83	3.74	0.8	−0.03	0.48	0.74	66	100	0.74	3.76	0.83	3.73	0.86	−0.03	0.59	0.8	73	100	0.8	3.76	0.83	3.75	0.81	−0.02	0.84	0.06	0.1
VB396487	461	3.61	0.78	3.64	0.8	0.03	0.54	0.75	71	100	0.76	3.61	0.78	3.62	0.86	0.01	0.6	0.8	74	100	0.8	3.61	0.78	3.62	0.81	0.01	0.83	0.04	0.08
VB396488	471	3.61	0.82	3.55	0.83	−0.07	0.54	0.75	70	99	0.76	3.61	0.82	3.57	0.89	−0.04	0.58	0.78	72	99	0.79	3.61	0.82	3.58	0.83	−0.03	0.82	0.03	0.07
VB396489	458	3.69	0.85	3.67	0.84	−0.02	0.55	0.79	70	100	0.79	3.69	0.85	3.73	0.88	0.04	0.57	0.79	71	99	0.79	3.69	0.85	3.71	0.8	0.02	0.83	0	0.04
VB396490	469	3.61	0.87	3.57	0.85	−0.05	0.5	0.75	67	99	0.75	3.61	0.87	3.49	0.91	−0.14	0.55	0.81	70	100	0.82	3.61	0.87	3.51	0.87	−0.12	0.84	0.06	0.09
VB398183	460	3.6	0.88	3.53	0.81	−0.08	0.51	0.73	67	98	0.74	3.6	0.88	3.52	0.87	−0.09	0.51	0.76	67	99	0.76	3.6	0.88	3.52	0.83	−0.09	0.83	0.02	0.09
VB398184	450	3.69	0.81	3.68	0.85	−0.02	0.54	0.75	70	99	0.75	3.69	0.81	3.68	0.88	−0.01	0.62	0.79	74	99	0.8	3.69	0.81	3.68	0.84	−0.01	0.84	0.04	0.09
VB398185	450	3.72	0.82	3.65	0.74	−0.09	0.54	0.74	71	99	0.75	3.72	0.82	3.69	0.86	−0.04	0.59	0.8	72	100	0.8	3.72	0.82	3.67	0.79	−0.06	0.82	0.05	0.07
VB398186	454	3.68	0.84	3.76	0.84	0.09	0.54	0.76	69	99	0.76	3.68	0.84	3.71	0.89	0.03	0.58	0.8	72	99	0.8	3.68	0.84	3.69	0.83	0.01	0.84	0.04	0.07
VB398187	460	3.75	0.82	3.76	0.85	0.01	0.5	0.75	67	100	0.75	3.75	0.82	3.72	0.89	−0.03	0.56	0.79	71	100	0.8	3.75	0.82	3.75	0.84	0	0.83	0.04	0.07
VB398188	448	3.76	0.83	3.76	0.85	0	0.46	0.73	64	99	0.73	3.76	0.83	3.75	0.85	−0.02	0.53	0.76	69	99	0.76	3.76	0.83	3.75	0.8	−0.02	0.82	0.03	0.09
VB398190	462	3.64	0.82	3.6	0.79	−0.05	0.47	0.71	66	99	0.71	3.64	0.82	3.63	0.78	0	0.58	0.77	73	99	0.78	3.64	0.82	3.64	0.74	0	0.82	0.07	0.11
VB398192	467	3.77	0.84	3.76	0.85	−0.02	0.49	0.74	67	99	0.74	3.77	0.84	3.74	0.86	−0.04	0.56	0.77	70	99	0.77	3.77	0.84	3.73	0.81	−0.05	0.81	0.03	0.06
VB398193	460	3.74	0.87	3.72	0.84	−0.03	0.48	0.73	66	98	0.73	3.74	0.87	3.74	0.86	−0.01	0.53	0.78	68	100	0.78	3.74	0.87	3.73	0.83	−0.01	0.83	0.05	0.11
VB398194	457	3.83	0.84	3.75	0.8	−0.09	0.47	0.7	66	98	0.71	3.83	0.84	3.77	0.86	−0.06	0.54	0.77	69	99	0.77	3.83	0.84	3.8	0.79	−0.04	0.81	0.07	0.11

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1; H2 = Human 2; h1er_rnd = H1 by e-rater (rounded to integers); h1er = H1 by e-rater (unrounded); *K* = kappa; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; SMD = standardized mean difference.

Table B3 Subgroup Differences: Generic Model

		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		QWK	*r*
Subgroup	*N*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*K*	QWK	% agree	% adj agree	*M*	*SD*	*SMD*	*r*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Female	41,889	3.70	0.81	3.69	0.81	−0.01	0.51	0.74	68.22	99.15	0.74	3.70	0.85	0.56	0.77	71.09	99.40	3.70	0.80	0.00	0.82	0.04	0.08
Male	14,696	3.62	0.84	3.61	0.84	−0.01	0.52	0.76	67.97	99.32	0.76	3.58	0.90	0.56	0.79	69.99	99.55	3.57	0.85	−0.05	0.83	0.03	0.07
White	44,849	3.75	0.80	3.74	0.80	−0.01	0.51	0.74	68.54	99.25	0.74	3.73	0.85	0.57	0.78	71.56	99.54	3.73	0.80	−0.02	0.82	0.04	0.08
African American	6,476	3.33	0.76	3.33	0.76	0.00	0.48	**0.69**	67.39	99.07	**0.69**	3.32	0.86	0.51	0.74	68.38	99.32	3.31	0.81	−0.02	0.79	0.05	0.09
Hispanic	1,948	3.46	0.88	3.45	0.86	−0.01	0.49	0.74	65.20	98.61	0.74	3.51	0.89	0.52	0.76	66.89	98.56	3.50	0.85	0.05	0.80	0.01	0.05
American Indian	314	3.45	0.77	3.49	0.81	0.05	0.38	**0.64**	60.83	98.09	**0.64**	3.45	0.90	0.47	0.71	64.65	98.41	3.44	0.83	−0.02	0.77	0.07	0.12
Other	1,372	3.59	0.85	3.59	0.85	−0.01	0.52	0.76	67.86	99.34	0.76	3.59	0.86	0.53	0.77	68.59	99.42	3.59	0.82	0.00	0.82	0.01	0.05
Unknown	104	3.63	0.84	3.62	0.85	−0.01	0.48	0.73	65.38	99.04	0.73	3.64	0.87	0.59	0.81	73.08	100.00	3.67	0.82	0.05	0.84	0.08	0.10
English as first language	52,467	3.70	0.81	3.69	0.81	−0.01	0.51	0.74	68.40	99.23	0.74	3.68	0.86	0.56	0.78	71.21	99.53	3.68	0.81	−0.02	0.82	0.04	0.08
Other first language	4,145	3.42	0.88	3.42	0.88	0.00	0.48	0.75	65.14	98.75	0.75	3.50	0.90	0.50	0.75	65.65	98.26	3.49	0.86	0.08	0.79	0.00	0.04

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1 score; H2 = Human 2 score; *K* = kappa; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; SMD = standardized mean difference.

Table B4 Subgroup Differences Under Contributory Score Model at 1.5 Threshold

			Human score by e-rater contributory score									
			Human score		E-rater contributory score		Stats					
		N	*M*	*SD*	*M*	*SD*	*SMD*	QWK	% agree	% adj agree	*r*
Gender	Female	22,926	7.39	1.53	7.39	1.53	0.00	0.95	1	96	0.95
	Male	8,068	7.17	1.63	7.17	1.63	−0.02	0.95	0	96	0.95
Ethnicity	African American	3,527	6.65	1.51	6.65	1.51	−0.01	0.93	1	94	0.94
	American Indian	183	7.06	1.54	7.06	1.54	−0.03	0.95	2	98	0.95
	Asian Pacific Islander	832	7.13	1.64	7.13	1.64	0.05	0.95	2	95	0.95
	Hispanic	1,056	6.94	1.66	6.94	1.66	0.02	0.95	1	93	0.95
	Other	745	7.16	1.56	7.16	1.56	0.01	0.95	0	95	0.95
	Unknown	62	7.29	1.53	7.29	1.53	0.03	0.94	0	94	0.94
	White	24,604	7.46	1.53	7.46	1.53	−0.01	0.95	0	97	0.95
First language	English	28,764	7.37	1.55	7.37	1.55	−0.01	0.95	0	96	0.95
	Other	2,245	6.90	1.67	6.90	1.67	0.03	0.95	2	93	0.95

Note. Adj = adjacent; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; *SMD* = standardized mean difference.

Table B5 Change in Scale Score Distribution Overall and by Gender and Ethnicity

	Difference scaled																	
	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10	11	Total
Overall	0	0.01	0.01	0.07	2.83	21.29	50.48	22.02	2.86	0.27	0.03	0.06	0.05	0.02	0	0	0	0	100
Male	0	0	0	0.03	0.88	5.92	13.46	5.36	0.72	0.07	0	0.02	0.02	0	0	0	0	0	26.47
Female	0	0.01	0.01	0.04	1.95	15.37	37.02	16.66	2.15	0.21	0.02	0.04	0.03	0.02	0	0	0	0	73.53
African American	0	0	0	0.02	0.41	2.46	5.8	2.43	0.4	0.04	0	0.01	0.01	0	0	0	0	0	11.58
American Indian	0	0	0	0	0.02	0.14	0.24	0.12	0.01	0	0	0	0	0	0	0	0	0	0.53
Asian Pacific Islander	0	0	0	0	0.06	0.46	1.33	0.74	0.11	0.02	0	0	0	0	0	0	0	0	2.72
Hispanic	0	0	0	0	0.11	0.65	1.7	0.78	0.16	0.02	0	0.01	0	0	0	0	0	0	3.43
Other	0	0	0	0	0.08	0.52	1.21	0.52	0.07	0.02	0	0	0	0	0	0	0	0	2.42
Unknown	0	0	0	0	0	0.04	0.09	0.03	0.01	0	0	0	0	0	0	0	0	0	0.18
White	0	0.01	0.01	0.05	2.15	17.04	40.11	17.4	2.09	0.18	0.02	0.04	0.03	0.01	0	0	0	0	79.13

Table B6 95% CIs for Simulated Scores by Gender and Ethnicity

				95% CI	
	N	*M* diff	SEM	Lower	Upper
Overall	65,883	0.02	0.02	174.69	174.76
Female	48,423	0.04	0.02	174.72	174.80
Male	17,428	−0.02	0.04	174.57	174.72
African American	7,631	0.01	0.05	171.78	171.97
American Indian	351	−0.07	0.25	172.63	173.60
Asian Pacific Islander	1,791	0.18	0.12	173.44	173.92
Hispanic	2,261	0.10	0.10	172.79	173.20
White	52,135	0.01	0.02	175.24	175.32

Note. CI = confidence interval; *M* diff = mean difference (simulated-operational); SEM = standard error of measurement.

Table B7 Pass Rate Overall and by Gender and Ethnicity

	N	Variable	*N*	Proportion passing	*SD*	*M*
Overall current	65,883	P_F_op	61,423	0.68	0.47	174.7
Overall simulation		P_F		0.68	0.47	174.7
Female	48,423	Current rate	45,102	0.68	0.47	174.8
		Simulated rate		0.69	0.46	174.8
Male	17,428	Current rate	16,290	0.66	0.47	174.6
		Simulated rate		0.67	0.47	174.6
African American	7,631	Current rate	6,884	0.42	0.49	171.9
		Simulated rate		0.43	0.50	171.9
American Indian	351	Current rate	302	0.5	0.50	173.1
		Simulated rate		0.51	0.50	173.1
Asian Pacific Islander	1,791	Current rate	1,584	0.64	0.48	173.7
		Simulated rate		0.66	0.47	173.9
Hispanic	2,261	Current rate	1,977	0.55	0.50	173
		Simulated rate		0.56	0.50	173.1
White	52,135	Current rate	49,139	0.72	0.45	175.3
		Simulated rate		0.73	0.45	175.3

Table B8 Pass Rate by State

Passing score	DI_STATE	*N*	Variable	*N*	Proportion passing	*SD*	*M*
174	AK	414	Current rate	414	0.68	0.47	176
			Simulated rate	414	0.7	0.46	
173	AL	76	Current rate	76	0.62	0.49	174
			Simulated rate	76	0.59	0.5	
173	AR	3,232	Current rate	3,232	0.65	0.48	175
			Simulated rate	3,232	0.65	0.48	
173	AZ	424	Current rate	424	0.55	0.5	173
			Simulated rate	424	0.55	0.5	
173	CA	97	Current rate	97	0.87	0.34	177
			Simulated rate	97	0.86	0.35	
173	CO	363	Current rate	363	0.78	0.42	176
			Simulated rate	363	0.79	0.41	
171	CT	1,376	Current rate	1,376	0.85	0.36	175
			Simulated rate	1,376	0.86	0.35	
171	DC	908	Current rate	908	0.9	0.31	178
			Simulated rate	908	0.9	0.3	
173	DE	850	Current rate	850	0.69	0.46	175
			Simulated rate	850	0.69	0.46	
173	FL	28	Current rate	28	0.57	0.5	175
			Simulated rate	28	0.57	0.5	
173	GA	45	Current rate	45	0.44	0.5	173
			Simulated rate	45	0.44	0.5	
170	GU	145	Current rate	145	0.77	0.42	173
			Simulated rate	145	0.79	0.41	
171	HI	645	Current rate	645	0.83	0.38	176
			Simulated rate	645	0.84	0.37	
173	IA	1,629	Current rate	1,629	0.75	0.44	175
			Simulated rate	1,629	0.75	0.44	

Table B8 (continued)

Passing score	DI_STATE	*N*	Variable	*N*	Proportion passing	*SD*	*M*
173	ID	447	Current rate	447	0.73	0.44	175
			Simulated rate	447	0.73	0.44	
173	IL	55	Current rate	55	0.64	0.49	175
			Simulated rate	55	0.66	0.48	
172	IN	4,621	Current rate	4,621	0.83	0.37	176
			Simulated rate	4,621	0.83	0.37	
173	KS	918	Current rate	918	0.67	0.47	175
			Simulated rate	918	0.67	0.47	
173	KY	415	Current rate	415	0.49	0.5	173
			Simulated rate	415	0.5	0.5	
175	LA	2,419	Current rate	2,419	0.28	0.45	172
			Simulated rate	2,419	0.28	0.45	
173	MA	49	Current rate	49	0.67	0.47	176
			Simulated rate	49	0.67	0.47	
173	MD	2,775	Current rate	2,775	0.69	0.46	175
			Simulated rate	2,775	0.7	0.46	
172	ME	761	Current rate	761	0.86	0.35	176
			Simulated rate	761	0.85	0.36	
173	MI	146	Current rate	146	0.59	0.49	173
			Simulated rate	146	0.56	0.5	
173	MN	2,753	Current rate	2,753	0.73	0.44	176
			Simulated rate	2,753	0.74	0.44	
173	MO	40	Current rate	40	0.73	0.45	175
			Simulated rate	40	0.75	0.44	
172	MS	3,030	Current rate	3,030	0.53	0.5	172
			Simulated rate	3,030	0.53	0.5	
173	MT	53	Current rate	53	0.81	0.4	177
			Simulated rate	53	0.87	0.34	
173	NC	2,827	Current rate	2,827	0.6	0.49	174
			Simulated rate	2,827	0.61	0.49	
173	ND	523	Current rate	523	0.65	0.48	174
			Simulated rate	523	0.65	0.48	
172	NE	1,816	Current rate	1,816	0.74	0.44	175
			Simulated rate	1,816	0.74	0.44	
172	NH	543	Current rate	543	0.81	0.39	176
			Simulated rate	543	0.8	0.4	
173	NJ	1,506	Current rate	1,506	0.7	0.46	175
			Simulated rate	1,506	0.71	0.45	
173	NM	4	Current rate	4	1	0	181
			Simulated rate	4	1	0	
172	NV	1,189	Current rate	1,189	0.68	0.47	174
			Simulated rate	1,189	0.69	0.46	
173	NY	118	Current rate	118	0.81	0.39	178
			Simulated rate	118	0.82	0.38	
173	OH	1,815	Current rate	1,815	0.59	0.49	174
			Simulated rate	1,815	0.6	0.49	
172	OK	17	Current rate	17	0.77	0.44	176
			Simulated rate	17	0.77	0.44	
171	OR	490	Current rate	490	0.94	0.24	178
			Simulated rate	490	0.94	0.24	
173	PA	7,961	Current rate	7,961	0.78	0.42	176
			Simulated rate	7,961	0.78	0.42	
173	RI	502	Current rate	502	0.73	0.44	175
			Simulated rate	502	0.74	0.44	

Table B8 (continued)

Passing score	DI_STATE	*N*	Variable	*N*	Proportion passing	*SD*	*M*
173	SC	2,026	Current rate	2,026	0.53	0.5	173
			Simulated rate	2,026	0.54	0.5	
173	SD	270	Current rate	270	0.68	0.47	175
			Simulated rate	270	0.68	0.47	
173	TN	2,470	Current rate	2,470	0.57	0.5	174
			Simulated rate	2,470	0.58	0.49	
173	TX	34	Current rate	34	0.56	0.5	174
			Simulated rate	34	0.53	0.51	
173	UT	955	Current rate	955	0.76	0.43	175
			Simulated rate	955	0.76	0.43	
176	VA	2,410	Current rate	2,410	0.45	0.5	175
			Simulated rate	2,410	0.46	0.5	
172	VI	18	Current rate	18	0.44	0.51	171
			Simulated rate	18	0.5	0.51	
174	VT	290	Current rate	290	0.68	0.47	175
			Simulated rate	290	0.7	0.46	
173	WA	73	Current rate	73	0.82	0.39	176
			Simulated rate	73	0.8	0.41	
174	WI	3,928	Current rate	3,928	0.7	0.46	176
			Simulated rate	3,928	0.71	0.46	
172	WV	912	Current rate	912	0.68	0.47	174
			Simulated rate	912	0.68	0.47	
173	WY	12	Current rate	12	0.75	0.45	176
			Simulated rate	12	0.75	0.45	

Note. DI = designated institute.

Table B9 Agreement With Human Scores at Prompt Level: Updated Generic Model

		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		QWK	*r*
Prompt	*N*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*K*	QWK	% agree	% adj agree	*M*	*SD*	*SMD*	*r*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
VB315223	479	3.68	0.85	3.64	0.82	−0.04	0.49	0.73	66	99	0.73	3.64	0.84	0.55	0.78	70	100	3.63	0.79	−0.06	0.83	0.06	0.10
VB315224	482	3.73	0.79	3.75	0.83	0.03	0.51	0.74	68	99	0.74	3.72	0.86	0.58	0.78	72	99	3.70	0.82	−0.04	0.82	0.04	0.07
VB315225	476	3.74	0.80	3.71	0.78	−0.04	0.56	0.75	72	99	0.75	3.70	0.86	0.55	0.78	71	100	3.70	0.79	−0.06	0.82	0.03	0.07
VB315226	630	3.68	0.85	3.69	0.85	0.00	0.55	0.77	70	99	0.77	3.73	0.88	0.56	0.80	71	100	3.71	0.84	0.03	0.84	0.02	0.06
VB315227	578	3.68	0.80	3.73	0.85	0.06	0.49	0.73	66	99	0.73	3.78	0.89	0.56	0.77	70	99	3.76	0.83	0.10	0.81	0.05	0.09
VB315228	478	3.57	0.77	3.56	0.78	−0.01	0.47	0.70	66	99	0.70	3.51	0.86	0.54	0.77	70	100	3.53	0.80	−0.05	0.80	0.07	0.09
VB315231	480	3.71	0.82	3.73	0.80	0.03	0.53	0.76	70	100	0.76	3.76	0.85	0.58	0.79	72	99	3.74	0.80	0.04	0.82	0.03	0.06
VB315232	654	3.54	0.79	3.52	0.83	−0.02	0.46	0.71	65	99	0.71	3.47	0.87	0.49	0.75	66	100	3.47	0.83	−0.09	0.79	0.04	0.09
VB315233	471	3.64	0.79	3.69	0.79	0.07	0.49	0.73	68	99	0.73	3.61	0.83	0.54	0.76	70	100	3.59	0.78	−0.07	0.82	0.04	0.09
VB315234	468	3.70	0.82	3.70	0.80	0.00	0.50	0.74	67	100	0.74	3.69	0.86	0.51	0.77	68	100	3.68	0.81	−0.02	0.81	0.03	0.07
VB315235	470	3.60	0.81	3.57	0.85	−0.04	0.51	0.74	68	99	0.74	3.50	0.87	0.57	0.80	72	100	3.48	0.83	**−0.16**	0.84	0.06	0.09
VB384839	461	3.64	0.79	3.71	0.80	0.09	0.51	0.73	69	99	0.73	3.67	0.84	0.51	0.75	68	100	3.67	0.79	0.04	0.80	0.02	0.07
VB384840	476	3.63	0.84	3.63	0.82	0.00	0.53	0.75	70	99	0.75	3.58	0.89	0.55	0.79	70	100	3.57	0.84	−0.08	0.83	0.04	0.08
VB384841	658	3.58	0.86	3.57	0.81	−0.01	0.51	0.76	68	100	0.76	3.55	0.87	0.55	0.79	70	100	3.55	0.84	−0.03	0.82	0.03	0.06
VB396482	486	3.78	0.79	3.81	0.79	0.04	0.50	0.72	68	99	0.72	3.79	0.86	0.52	0.75	68	99	3.78	0.80	−0.01	0.80	0.04	0.08
VB396483	483	3.65	0.75	3.65	0.76	0.01	0.51	0.74	70	100	0.74	3.64	0.85	0.55	0.77	71	100	3.63	0.80	−0.03	0.82	0.04	0.08
VB396484	477	3.56	0.82	3.55	0.82	−0.02	0.54	0.76	70	99	0.76	3.51	0.86	0.54	0.78	70	100	3.51	0.83	−0.06	0.83	0.02	0.06
VB396485	656	3.62	0.91	3.63	0.86	0.00	0.45	0.74	63	99	0.74	3.69	0.92	0.53	0.79	67	99	3.69	0.87	0.07	0.82	0.05	0.09
VB396486	481	3.73	0.85	3.72	0.82	−0.01	0.49	0.75	66	100	0.75	3.70	0.90	0.58	0.81	72	100	3.70	0.85	−0.04	0.84	0.05	0.09
VB396487	666	3.61	0.79	3.64	0.81	0.05	0.52	0.73	69	100	0.73	3.58	0.90	0.56	0.78	71	100	3.59	0.84	−0.02	0.81	0.05	0.08
VB396488	481	3.62	0.82	3.56	0.81	−0.07	0.54	0.75	70	99	0.75	3.52	0.87	0.55	0.76	70	99	3.53	0.83	−0.10	0.81	0.02	0.06
VB396489	464	3.65	0.86	3.64	0.85	−0.01	0.55	0.79	70	100	0.79	3.63	0.88	0.56	0.79	70	99	3.62	0.82	−0.04	0.84	0.00	0.05
VB396490	481	3.62	0.88	3.58	0.86	−0.04	0.51	0.76	67	99	0.76	3.50	0.90	0.54	0.79	69	99	3.51	0.87	−0.13	0.85	0.03	0.08
VB398183	659	3.61	0.89	3.56	0.85	−0.06	0.54	0.77	69	98	0.77	3.57	0.89	0.55	0.78	70	99	3.58	0.86	−0.04	0.84	0.02	0.07
VB398184	475	3.67	0.81	3.66	0.86	−0.02	0.52	0.75	68	99	0.75	3.67	0.87	0.60	0.80	73	100	3.67	0.83	0.00	0.84	0.05	0.09
VB398185	593	3.70	0.82	3.63	0.77	−0.09	0.58	0.77	73	99	0.78	3.72	0.88	0.57	0.80	71	100	3.71	0.82	0.01	0.83	0.02	0.06
VB398186	464	3.64	0.86	3.70	0.85	0.07	0.56	0.78	70	99	0.78	3.61	0.90	0.54	0.79	69	99	3.60	0.85	−0.05	0.84	0.01	0.06
VB398187	477	3.71	0.85	3.73	0.86	0.02	0.51	0.77	67	100	0.77	3.70	0.89	0.60	0.81	73	100	3.69	0.85	−0.02	0.84	0.05	0.08
VB398188	463	3.76	0.84	3.75	0.86	0.00	0.47	0.74	65	99	0.74	3.75	0.87	0.56	0.78	70	99	3.74	0.83	−0.02	0.84	0.05	0.10
VB398190	671	3.63	0.85	3.60	0.83	−0.03	0.50	0.74	68	99	0.74	3.66	0.84	0.55	0.78	70	99	3.65	0.81	0.03	0.82	0.04	0.08
VB398192	468	3.75	0.87	3.74	0.88	−0.01	0.50	0.77	67	99	0.77	3.66	0.87	0.52	0.77	68	99	3.66	0.82	−0.10	0.81	0.00	0.05
VB398193	479	3.72	0.88	3.70	0.86	−0.02	0.49	0.74	66	99	0.74	3.74	0.87	0.58	0.81	72	100	3.73	0.84	0.01	0.84	0.07	0.10
VB398194	478	3.81	0.84	3.73	0.82	−0.09	0.48	0.72	66	98	0.72	3.75	0.84	0.53	0.76	69	99	3.77	0.79	−0.05	0.82	0.05	0.10

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1 score; H2 = Human2 score; *K* = kappa; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; *SMD* = standardized mean difference.

Table B10 Subgroup Differences: Updated Generic Model

		H1 by H2										H1 by e-rater (rounded to integers)						H1 by e-rater (unrounded)				Degradation	
		H1		H2		Stats						e-rater						e-rater		Stats		QWK	*r*
Subgroup	*N*	*M*	*SD*	*M*	*SD*	*SMD*	*K*	QWK	% agree	% adj agree	*r*	*M*	*SD*	*K*	QWK	% agree	% adj agree	*M*	*SD*	*SMD*	*r*	H1 by e-rater rounded – H1 by H2	H1 by e-rater unrounded – H1 by H2
Female	11,405	3.69	0.83	3.68	0.82	0.00	0.51	0.74	68.18	99.07	0.74	3.67	0.86	0.55	0.78	70.23	99.47	3.66	0.81	−0.03	0.82	0.04	0.08
Male	4,222	3.59	0.85	3.58	0.85	−0.01	0.51	0.75	67.43	99.10	0.75	3.54	0.90	0.55	0.79	69.85	99.60	3.53	0.86	−0.07	0.83	0.04	0.08
White	12,354	3.74	0.82	3.73	0.81	0.00	0.51	0.74	68.41	99.14	0.74	3.71	0.85	0.56	0.78	71.09	99.57	3.70	0.80	−0.05	0.82	0.04	0.08
African America	1,819	3.29	0.79	3.29	0.79	0.00	0.49	0.72	67.40	99.23	0.72	3.23	0.88	0.50	0.75	66.79	99.34	3.24	0.84	−0.07	0.80	0.03	0.08
Hispanic	535	3.45	0.88	3.44	0.88	−0.01	0.53	0.75	67.85	97.94	0.75	3.45	0.89	0.53	0.77	67.66	98.50	3.46	0.85	0.01	0.80	0.01	0.05
Asian Pacific Islander	453	3.49	0.88	3.45	0.88	−0.05	0.45	0.73	62.91	98.68	0.73	3.59	0.91	0.48	0.77	64.46	99.34	3.58	0.86	0.10	0.81	0.04	0.08
American Indian	69	3.39	0.69	3.43	0.78	0.06	0.41	**0.63**	65.22	98.55	**0.64**	3.26	0.80	0.44	0.70	66.67	100.00	3.29	0.79	−0.14	0.76	0.07	0.12
Other	380	3.56	0.92	3.54	0.90	−0.02	0.47	0.75	63.68	98.42	0.75	3.47	0.90	0.50	0.78	65.53	99.47	3.49	0.85	−0.07	0.82	0.03	0.07
English as first language	14,450	3.69	0.82	3.68	0.82	−0.01	0.51	0.74	68.10	99.14	0.74	3.65	0.87	0.56	0.79	70.55	99.61	3.64	0.82	−0.05	0.83	0.04	0.08
Other first language	699	3.55	0.89	3.56	0.88	0.01	0.55	0.77	69.24	98.71	0.77	3.57	0.92	0.51	0.77	66.38	98.43	3.57	0.88	0.02	0.81	−0.01	0.04

Note. Values that failed to meet the recommended threshold for the statistic are in bold. Adj = adjacent; H1 = Human 1 score; H2 = Human2 score; *K* = kappa; *M* = mean; QWK = quadratic-weighted kappa; *r* = correlation; *SD* = standard deviation; *SMD* = standardized mean difference.

Suggested citation:

Ramineni, C., Trapani, C. S., & Williamson, D. M. (2014). *Evaluation of e-rater® for the Praxis I® writing test* (ETS Research Report No. RR-15-03). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12047

Action Editor: James Carlson

Reviewers: Jay F. Breyer and Brent Bridgeman